SD-TSIA211

Correction - TD2

2018

Exercise 9 : (Proximal stochastic gradient for logistic regression)

1. $\forall i \in \{1, ..., n\}, \mathbb{P}_{w,w_0}(Y_i = 1|x_i) = \frac{\exp\left(x_i^\top w + w_0\right)}{1+\exp\left(x_i^\top w + w_0\right)} = \frac{1}{1+\exp\left(-(x_i^\top w + w_0)\right)}$.

   $\mathbb{P}_{w,w_0}(Y_i = -1|x_i) = 1 - \mathbb{P}_{w,w_0}(Y_i = 1|x_i) = \frac{1}{1+\exp\left(x_i^\top w + w_0\right)}$.
   Hence, $\forall i \in \{1, ..., n\}, \mathbb{P}\left(Y_i = y_i|x_i\right) = \frac{1}{1+\exp\left(-y_i\left(x_i^\top w + w_0\right)\right)}$.
2. By independence of the observations $(x_i, y_i)_{1 \le i \le n}$, the likelihood writes :

   $\mathbb{P}\left(Y_1 = y_1, ..., Y_n = y_n|x_1, ..., x_n\right) = \prod\limits_{i=1}^{n} \mathbb{P}\left(Y_i = y_i|x_i\right)$.
   Then, one can compute the log-likelihood and use question 1.
   The maximum likelihood estimator minimizes the opposite of the log-likelihood :

   $$\hat{w} = \arg\min_{w} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i(x_i^\top w + w_0)\right)\right)$$

3. Denote $f_i(w, w_0) := \log\left(1 + \exp\left(-y_i(x_i^\top w + w_0)\right)\right)$.
   Observe that $f_i(w, w_0) = h(-y_i(x_i^\top w + w_0))$ where $h(u) := \log\left(1 + \exp(u)\right)$.
   Then, $\nabla f(w, w_0) = \sum\limits_{i=1}^{n} \nabla f_i(w, w_0)$ and by the chain rule, one has :

   $\nabla f_i(w, w_0) = \frac{-y_i}{1+\exp\left(-y_i\left(x_i^\top w + w_0\right)\right)} \begin{pmatrix} x_i \\ 1 \end{pmatrix}$.

   $$\nabla f(w, w_0) = \sum_{i=1}^{n} \frac{-y_i}{1 + \exp\left(-y_i\left(x_i^\top w + w_0\right)\right)} \begin{pmatrix} x_i \\ 1 \end{pmatrix}.$$

4. Denote $v(x) = \frac{\lambda}{2}\|x\|^2$.
   We derive its proximal operator from the application of Fermat's rule.
   $p := \text{prox}_v(x) = \underset{y}{\text{argmin}}\{v(y) + \frac{1}{2}\|y - x\|^2\}$.
   This is equivalent to : $0 = \nabla h(p) + z - p = (1 + \lambda)z - x$. Hence,

   $$\text{prox}_v(x) = \frac{1}{1 + \lambda}x.$$

Remark : In this case, the subdifferential of the sum is the sum of the subdifferentials given that $0 \in \mathrm{ri}(\mathrm{dom}v - \mathrm{dom}u) = \mathbb{R}^p$ where u is the second (quadratic) function in the argmin defining the proximal operator.

5. Notice that the two functions of the optimization problem are convex, closed (even continuous) and their domains are nonempty. The first one is differentiable and L-smooth (we do not explicit this constant here). The proximal stochastic gradient method writes as follows :
   - Draw i uniformly at random from $\{1, ..., n\}$.
   - Update rule : $w_{k+1} = \mathrm{prox}_v\,(w_k - \gamma \nabla f_i(w_k))$ and $w_{0,k+1} = w_{0,k} - \gamma \nabla f_i(w_{0,k})$ where $\nabla f_i(w_k)$ and $\nabla f_i(w_{0,k})$ are the components of the vector $\nabla f_i(w_k, w_{0,k})$. Questions 3 and 4 provide the expressions to write the iterations of the algorithm explicitly. γ is the stepsize of the algorithm (One can take $\gamma = \frac{1}{L}$).

Exercise 7 : (LASSO)

1. By Fermat's rule, 0 is a solution to the LASSO problem is equivalent to $0 \in -A^\top b + \lambda\partial(\|\cdot\|_1)(0)$. It has already been shown in the lecture that $\partial(\|\cdot\|_1)(0) = [-1, 1]^n$. The latter inclusion condition is satisfied for all $\lambda \geq \|A^\top b\|_\infty$. Thus, for large enough λ, $\{0\}$ is a solution and one can verify its uniqueness.
2. Recall that the proximal operator of the l^1 norm is the soft thresholding operator S (already seen in the lecture). The proximal gradient algorithm corresponds to the following update rule :

$$x_{k+1} = \mathrm{prox}_{\lambda\|\cdot\|_1}(x_k - \gamma \nabla f(x_k))$$

   where $f(x) := \frac{1}{2}\|Ax - b\|_2^2$. Define L the lipschitz constant corresponding to the gradient of f ($L = \|A^\top A\|$). We use the stepsize $\gamma = \frac{1}{L}$ as a stepsize. Therefore, the update rule writes :

$$x_{k+1} = S_\lambda\left(x_k - \frac{1}{L}A^\top(Ax_k - b)\right)$$

   where S_λ is the coordinatewise operator defined by $S_{\lambda,i}(x_i) = x_i - \lambda$ if $x_i > \lambda$, 0 if $|x_i| \leq \lambda$ and $x_i + \lambda$ if $x_i < -\lambda$ for $i \in \{1, ..., n\}$.
3. We recall the following convergence result for the proximal gradient algorithm (under convexity and stepsize assumptions precised in the lecture) :

$$(f + g)(x_k) - \inf(f + g) \leq \frac{LD^2}{2k}$$

   Setting $\frac{LD^2}{2k} \leq \epsilon$, we have that the number of iterations should verify $k \geq \frac{LD^2}{2\epsilon}$.

   Remark : a similar result can be derived for the iterates (for an ϵ-minimizer) with the rate of convergence of the iterates.

Exercise 10 : (Optimization with explicit constraints)

1. $f(x) + \iota_C(x) = f(x)$ if $x \in C$, and $+\infty$ otherwise. The result follows from this remark.
2. Recall the definition of the subdifferential :
$\partial\iota_C(x) = \{q \in \mathbb{R}^n : \forall y \in \mathrm{dom}\iota_C,\ \iota_C(y) \geq \iota_C(x) + \langle q, y - x\rangle\}$.
For $x \in C$, $\partial\iota_C(x) = \{q \in \mathbb{R}^n : \forall y \in C, \langle q, y - x\rangle \leq 0\}$ and for $x \notin C, \iota_C(x) = \emptyset$. Using the same definition, one can see that for any $x \in C$ and any $q \in \iota_C(x)$, $\lambda q \in \iota_C(x)$ for any $\lambda \geq 0$: $\iota_C(x)$ is a cone.
3. Using Fermat's rule, $x_* \in \underset{x\in\mathbb{R}^d}{\mathrm{argmin}}\,\{f(x) + \iota_C(x)\}$ is equivalent to $0 \in \nabla f(x_*) + \partial\iota_C(x_*)$. This concludes the proof.
Remark : Once again, in our case, the subdifferential of the sum is the sum of the subdifferentials since $0 \in \mathrm{ri}(\mathbb{R}^d - C) = \mathbb{R}^d$.
4. $\mathcal{H}_{w,b} = \{x \in \mathcal{X} : \langle w, x\rangle + b = 0\}$ is a convex set (use the definition). Using question 2, we find that : $\partial\iota_{\mathcal{H}_{w,b}}(x) = \{\lambda w, \lambda \in \mathbb{R}\}$ for $x \in \mathcal{H}_{w,b}$ and $\emptyset$ otherwise.
5. $\underset{x\in\mathcal{H}_{w,b}}{\mathrm{argmin}}\|x - z\|_2 = \underset{x\in\mathcal{H}_{w,b}}{\mathrm{argmin}}\frac{1}{2}\|x - z\|_2^2$. Applying the results of questions 3 and 4 to the latter optimization problem, we get that x^* is a solution if and only if $-(x^* - z) \in \partial\iota_{\mathcal{H}_{w,b}}(x) = \{\lambda w, \lambda \in \mathbb{R}\}$. Therefore, there exists $\lambda \in \mathbb{R}$ such that $x^* = -\lambda w + z$. Since $x^* \in \mathcal{H}_{w,b}$, $\langle w, x^*\rangle + b = 0$. We determine λ by substituting x^* in the hyperplan equation : $\lambda = \frac{\langle w,z\rangle + b}{\|w\|_2^2}$. Hence,

$$d\left(z, \mathcal{H}_{w,b}\right) = \min_{x\in\mathcal{H}_{w,b}} \|x - z\|_2 = \|x^* - z\|_2 = |\lambda|\|w\|_2 = \frac{|\langle w, z\rangle + b|}{\|w\|_2}$$